Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 1219–1224

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, 2019, Leuven, Belgium

Preservation of confidential information privacy and association rule hiding for data mining: a bibliometric review

Jesus Silva[a]*, Jenny Cubillos[b], Jesus Vargas Villa[c], Ligia Romero[d], Darwin Solano[e], [f]Claudia Fernández

[a] *Universidad Peruana de Ciencias Aplicadas, Lima 07001 , Peru*
[b] *Fundación Universitaria Konrad Lorenz, Bogota 110111, Colombia*
[c,d,f] *Universidad de la Costa (CUC), Barranquilla 080003, Colombia*

Abstract

In this era of technology, data of business organizations are growing with acceleration. Mining hidden patterns from this huge database would benefit many industries improving their decision-making processes. Along with the non-sensitive information, these databases also contain some sensitive information about customers. During the mining process, sensitive information about a person can get leaked, resulting in a misuse of the data and causing loss to an individual. The privacy preserving data mining can bring a solution to this problem, helping provide the benefits of mined data along with maintaining the privacy of the sensitive information. Hence, there is a growing interest in the scientific community for developing new approaches to hide the mined sensitive information. In this research, a bibliometric review is carried out during the period 2010 to 2018 to analyze the growth of studies regarding the confidential information privacy preservation through approaches addressed to the hiding of association rules of data.

Keywords: confidential information privacy preservation; approaches to hiding of association rules of data; bibliometric analysis; SCOPUS.

* Corresponding author. Tel.: +51920287620
E-mail address: jesussilvaUCP@gmail.com

10.1016/j.procs.2019.04.175

1. Introduction

Privacy Preserving Data Mining (PPDM) provides prevention of breach in the user´s privacy when the mined data is shared between multiple parties. Association Rule Hiding provides solution to the problem. In late nineties (Nguyen X, et al; 2012)[1], PPDM began and, in recent years, has produced great improvements in this issue since many techniques and approaches have been developed in this field of research (Doganay M. et al; 2008)[2], (Moustakides G and Verykios V; 2008)[3], (Adhvaryu R, Domadiya N; 2012)[4]. PPDM helps to take benefit from mined data while maintaining the secrecy of sensitive data and, at the same time, maintains the data accuracy. Nevertheless PPDM is a challenging task. There are many issues like privacy, trust, data quality (DQ), and malicious data mining intrusion detection systems that must be considered in the context of crucial online databases, thereby making this task more complex (Aggarwal C and Philip S; 2004)[5], (Moustakides G. and Verykios V; 2006)[6], (Bogdanov D et al; 2012)[7].

2. Method

This study was conducted based on the results obtained after a bibliometric review which helped identify the scope and investigations that have given greater relevance to the issue of privacy preservation of confidential information with an approach of association rule hiding from data mining in the period 2010 to 2018 (Dnyanesh P, et al; 2012)[8], (Li G and Wang Y; 2012)[9], (Li G and Xi M; 2015)[10], (Domadiya N and Rao U; 2013)[11], (Gaitán-Angulo M. et al; 2018)[12], (Lis-Gutiérrez J. et al; 2018)[13]. The search for information was carried out in the Scopus database and was guided under the following routes:

(TITLE-ABS-KEY (preservation AND OF AND CONFIDENTIAL AND INFORMATION) AND TITLE-ABS-KEY (data and mining)) AND PUBYEAR > 2009

(TITLE-ABS-KEY (rule and hiding AND approaches) AND TITLE-ABS-KEY (data AND association) AND TITLE-ABS-KEY (data and mining))

3. Results and Discussions

3.1 Preservation of the privacy of confidential information

As a result, 22 research articles were obtained, including terms related to data mining and preservation of the privacy of confidential information. The results are described below. Fig. 1 shows the countries that most publications have produced over the last 10 years, with India as the country with the highest number of articles with a total of 11 publications, followed by Spain with three articles, and Chile, Germany, and the United States with 2 publications each.

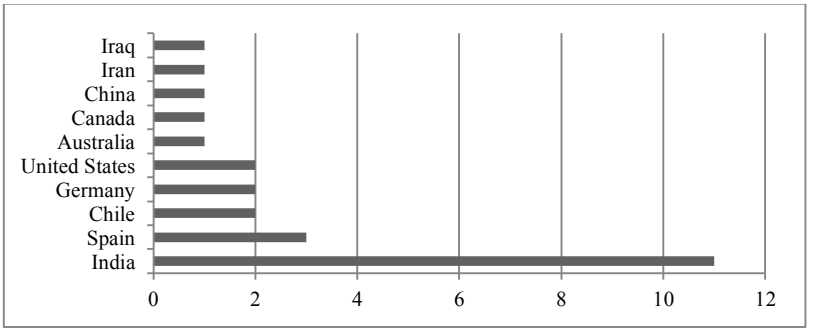

Fig 1. *Analysis of documents by country on the preservation of privacy policy*

Fig. 2 shows the fluctuation of the subject over the past few years. It can be observed that the year 2016 was the most productive period with 5 articles, and in 2017 the number of publications declined to 2. However, for the year 2018 publications continue to rise.

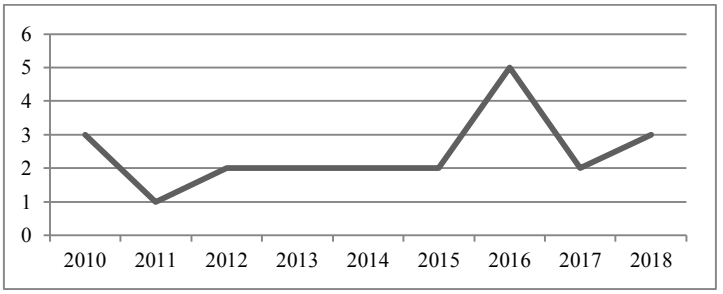

Fig 2. *Analysis of documents by year of publication on preservation of privacy policy*

In addition, an analysis of authors and keywords conducted by means of the bibliometric Vosviewer network visualization allows to make a visual analysis of the results obtained from the search performed, using clusters identified by color and size in the network. The above is shown in Fig 3.

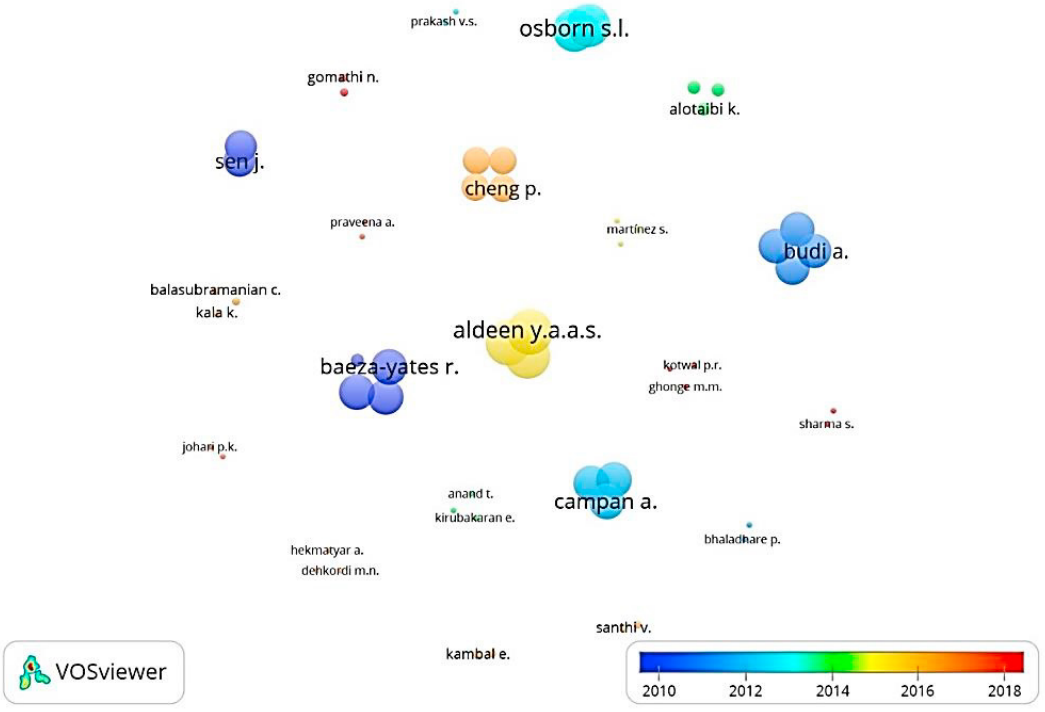

Fig 3. *Networks of authors on preservation of privacy policy*

Fig. 3 shows the networks of authors who have published on data mining and preservation of the privacy of confidential information, and the year in which the authors published them. For example Campan.A and its network of co-authors conducted publications during the year 2012, indicated by the color of the cluster where it is located. Within the most important authors is Osborn S.L, Budi A., and ALDEEN Y.A.A.S. However, these authors have published in different years and, therefore, do not have a network among themselves. The results obtained with respect to the keywords (Fig. 4) of the search indicate that there are three fourths interrelated terms, firstly, in red are

terms as concept, classification, publication, server, and clustering. Secondly, in dark green color are company, person, facebook, file, great interest, time, among other. This network is related to some terms of the network identified in blue color where are terms such as processes of anonymization, purpose, context, and order. And finally, the network in clear green with terms such as anonymity, system, and correlation.

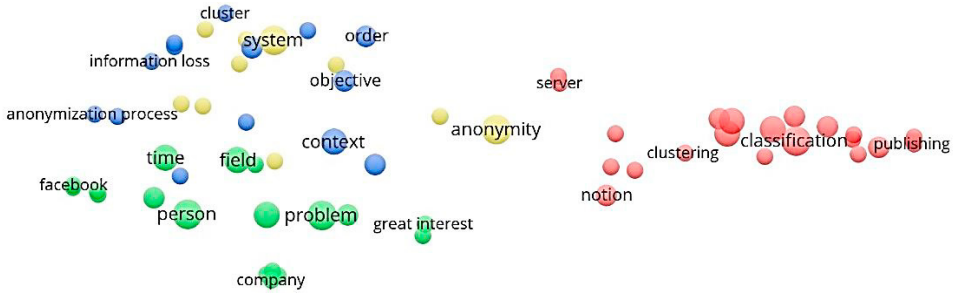

Fig 4. *Networks of keywords on the preservation of privacy policy*

3.2 Approaches to hiding of data association rules in data mining

Fig. 5 shows the countries that most publications have produced in recent years, with India as the country with the highest number of articles with 25 publications, followed by Taiwan with 21, China with 16, and the United States with 9. This means that these are the countries with more researches on the subject.

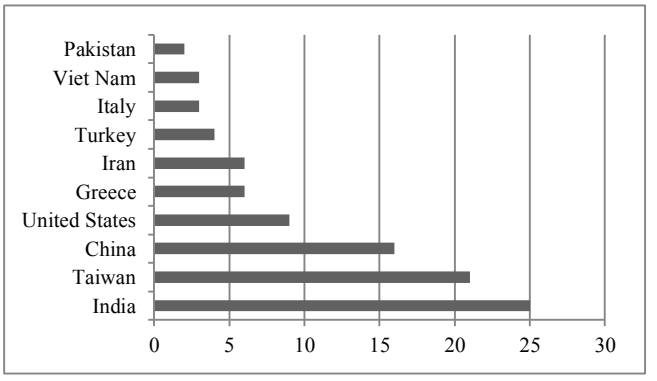

Fig 5. *Analysis of documents by country on rules of association of data in data mining*

Fig 6 shows the variation of the subject over the years. This figure evidences that the years 2011 and 2016 were the periods when more publications were carried out, while in the year 2018 the number of publications declined.

In addition, an analysis of authors and keywords was conducted (see fig 7) by means of the bibliometric Vosviewer network visualization which allows to make a visual analysis of the results obtained from the search performed, identifying clusters by color and size in the network.

Fig 7 shows the networks of authors who have published on approaches about hiding of association rules of data using data mining and the authors with whom the research networks are formed, which are identified by location and colors. Each network has the same color according to the authors who present relationship. A central author is Hong T.-P. who is related to the network of authors Huang J.- P., Lan G.-C., and Xul. Another important network is formed by Wang S.-L., Jafari a., Parikh B., and maskey R.

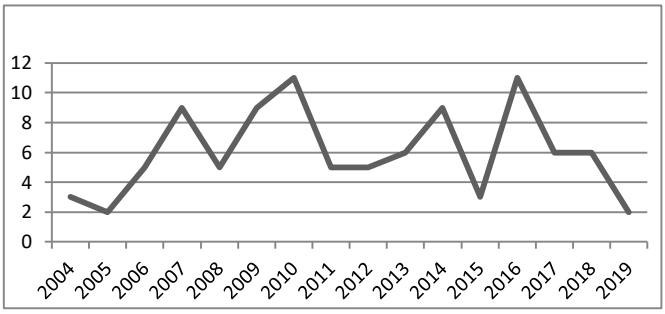

Fig 6. *Analysis of documents by year of publication on data association rules in data mining*

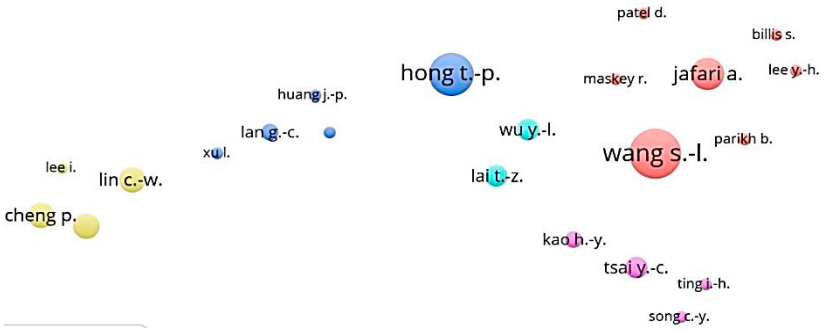

Fig 7. *Networks of authors on rules of association of data in data mining*

The results obtained with respect to the keywords for the search (see Fig 8) indicate that there are three fourths of networks of related terms between them. Firstly, in red are terms such as mining, knowledge, experiment, proceeding, organization, utility, impact, and research. Secondly, in dark green are frequent itemset, unauthorized access, sensitive rule, new algorithm, among others. The light green is the network related to some terms such as context, new approach, and kind. And finally, the network in blue with terms such as process, efficiency, data mining algorithm, process and confidentiality.

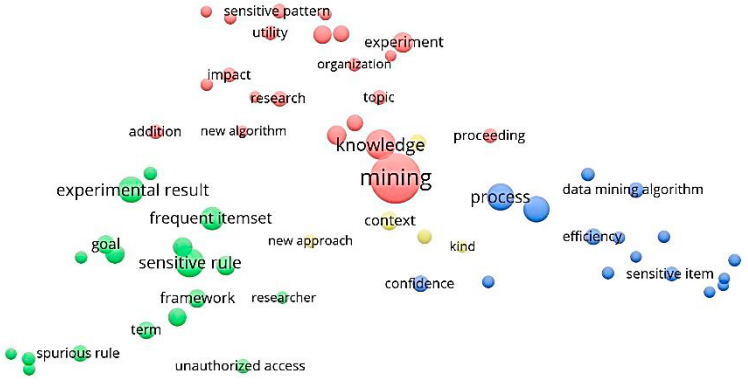

Fig 8. *Networks of keywords on association rules of data using data mining*

4. Conclusions

The objective of this paper is to make the researcher familiar with the current state-of-art and the future scope of these techniques. The literature survey carried out in this paper provides researchers a better understanding about the growth of this field and the issues related to PPDM. The authors evaluated different association rule hiding algorithms based upon various parameters like efficiency, scalability, privacy level, hiding failure, and quality of data. PPDM is applicable to different Data Mining fields like classification, clustering, association rule hiding, etc. The focus is emphasized on PPDM for association rule hiding developed in the latest decade.

References

[1] Nguyen XC, Le HB, Cao TA (2012). An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases. In: Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF) IEEE, pp: 1-4.

[2] Doganay MC, Pedersen TB, Saygin Y, Savaş E, Levi A (2008). Distributed privacy preserving k-means clustering with additive secret sharing. In: Proceedings of the 2008 international workshop on Privacy and anonymity in information society ACM, pp: 3-11.

[3] Moustakides G V and Verykios V S (2008). A maxmin approach for hiding frequent itemsets. Data and Knowledge Engineering 65(1):75–89.

[4] Adhvaryu R, Domadiya N (2012). An Improved EMHS Algorithm for Privacy Preserving in Association Rule Mining on Horizontally Partitioned Database. In: Security in Computing and Communications Springer Berlin Heidelberg, pp: 272-280.

[5] Aggarwal CC, Philip SY (2004). A condensation approach to privacy preserving data mining. In: Advances in Database Technology-EDBT Springer Berlin Heidelberg, pp. 183-199.

[6] Moustakides G V and Verykios V S (2006). A max–min approach for hiding frequent itemsets. In: Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), pp: 502–506.

[7] Bogdanov D, Talviste R, Willemson J (2012). Deploying secure multi-party computation for financial data analysis. In: Financial Cryptography and Data Security Springer Berlin Heidelberg, pp: 57-64.

[8] Dnyanesh P, Akhtar WS, Loknath S, TN R (2012). Perturbation Based Reliability And Maintaining Authentication In Data Mining. In: International Conference on Advances in Computer and Electrical Engineering, pp: 59-63.

[9] Li G, Wang Y (2012). A Privacy-Preserving Classification Method Based on Singular Value Decomposition. In: Int. Arab J. Inf. Technol.: 9(6):529-34.

[10] Li G, Xi M (2015). An Improved Algorithm for Privacy-preserving Data Mining Based on NMF. In: Journal of Information & Computational Science, 12(9), pp: 3423–3430.

[11] Domadiya NH and Rao UP (2013). Hiding sensitive association rules to maintain privacy and data quality in database. In: Advance Computing Conference, IEEE, pp: 1306-1310.

[12] Gaitán-Angulo M., Cubillos Díaz J., Viloria A., Lis-Gutiérrez JP., Rodríguez-Garnica P.A. (2018) Bibliometric Analysis of Social Innovation and Complexity (Databases Scopus and Dialnet 2007–2017). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[13] Lis-Gutiérrez J.P., Henao C., Zerda Á., Gaitán M., Correa J.C., Viloria A. (2018) Determinants of the Impact Factor of Publications: A Panel Model for Journals Indexed in Scopus 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham